

AD-A110 311

WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER

F/6 12/1

A GENERALIZED CONJUGATE GRADIENT METHOD FOR NON-SYMMETRIC SYSTEMS

ETC(U)

OCT 81 J C STRIKWERDA

DAAG29-80-C-0041

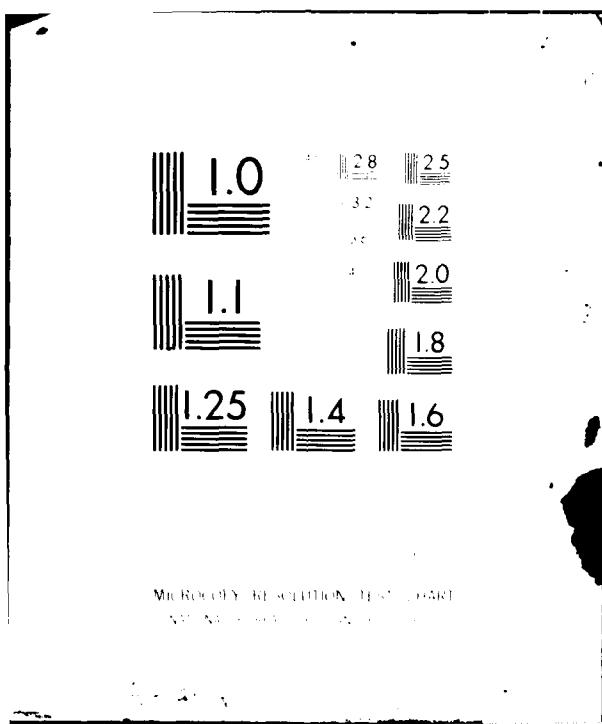
UNCLASSIFIED

MRC-TSR-2290

NL

REF:  
AD  
A.O.U.

END  
FILED  
02-82  
DTIC



MICRODOT RESOLUTION TEST CHART  
MITSUBISHI ELECTRIC

MRC FILE COPY

AD-A110311

LEVEL

(2)

MRC Technical Summary Report #2290

A GENERALIZED CONJUGATE GRADIENT  
METHOD FOR NON-SYMMETRIC SYSTEMS  
OF LINEAR EQUATIONS

John C. Strikwerda



**Mathematics Research Center  
University of Wisconsin-Madison  
610 Walnut Street  
Madison, Wisconsin 53706**

October 1981

(Received August 27, 1981)

Approved for public release  
Distribution unlimited

Sponsored by

U. S. Army Research Office  
P. O. Box 12211  
Research Triangle Park  
North Carolina 27709

National Science Foundation  
Washington, DC 20550

National Aeronautics and  
Space Administration  
Washington, DC 20546

231500  
82 02 03 079

UNIVERSITY OF WISCONSIN - MADISON  
MATHEMATICS RESEARCH CENTER

A GENERALIZED CONJUGATE GRADIENT METHOD  
FOR NON-SYMMETRIC SYSTEMS OF LINEAR EQUATIONS

John C. Strikwerda

Technical Summary Report #2290

October 1981

ABSTRACT

A new iterative method is presented for solving non-symmetric linear systems of equations. The method requires that the symmetric part of the matrix of the linear system be positive definite, and the method is efficient only if the symmetric part is easily invertible. The method is modeled on the conjugate gradient method for symmetric positive definite systems and has the finite termination property. The results from several numerical experiments are presented and compared with a similar method proposed by Concus, Golub, and Widlund.

AMS (MOS) Subject Classification: 65F10

Key Words: Conjugate-Gradient Method, Non-symmetric Systems

Work Unit Number 3 - Numerical Analysis and Computer Science

---

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041. This material is based upon work supported by the National Science Foundation under Grant No. MCS-7927062. Portions of this research were performed under NASA Contract Nos. NASI-15810 and NASI-16394 while the author was in residence at ICASE, NASA Langley Research Center, Hampton, VA 23665.

## SIGNIFICANCE AND EXPLANATION

In this report a new method is presented for the solution of linear systems of equations,

$$Ax = b ,$$

where the matrix  $A$  is a non-symmetric matrix. The matrix  $A$  can be written as the sum of its symmetric and skew-symmetric parts

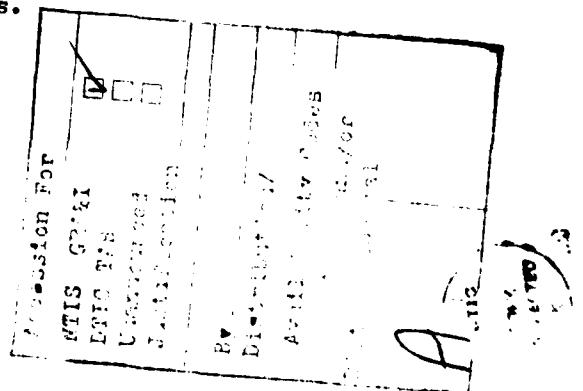
$$A = P + Q$$

and the method requires that the symmetric part,  $P$ , be positive definite, i.e.

$$(x, Px) > 0$$

for all non-zero vectors  $x$ . The method is efficient only when the solution of the linear system  $Py = c$  can be obtained easily. This is the case in many problems such as the solution of elliptic equations whose highest order part is the Laplacian.

The method is modeled on the conjugate gradient method which is a widely used method for symmetric positive definite systems.



The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.

A GENERALIZED CONJUGATE GRADIENT METHOD  
FOR NON-SYMMETRIC SYSTEMS OF LINEAR EQUATIONS

John C. Strikwerda

1. Introduction

In this paper a generalized conjugate gradient method for solving linear systems of equations is presented. The proto-type for this method is the system

$$(1.1) \quad (I + S)x = b$$

where  $S$  is a skew-symmetric matrix and  $I$  is the identity matrix. The method is derived as an acceleration of the steepest descent method for the system (1.1) where the norm of the residual is minimized.

Concus and Golub (1976) and Widlund (1978) have presented and discussed a generalized conjugate gradient method for non-symmetric linear systems which has some similarities with the present method (see also Hageman et al. 1980). Although both methods can be derived in several ways, the derivation of the method of this paper is different in spirit from the method of Concus and Golub (1976) and Widlund (1978) as given in their papers. Their method is derived by imposing orthogonality constraints on an appropriate sequence of vectors and the convergence properties are then deduced. The method presented here is derived as an acceleration of a steepest descent method and the orthogonality results then follow. As shown in section 4 the two methods have similar rates of convergence and in the numerical experiments they behaved similarly.

The method of this paper can be applied in Hilbert spaces as was done by Widlund (1978), but little would be gained by the extra generality so we will consider the method only for finite-dimensional spaces.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041. This material is based upon work supported by the National Science Foundation under Grant No. MCS-7927062. Portions of this research were performed under NASA Contract Nos. NASI-15810 and NASI-16394 while the author was in residence at ICASE, NASA Langley Research Center, Hampton, VA 23665.

## 2. Derivation of the Method.

Consider the system of linear equations

$$(2.1) \quad Ax = b$$

where  $A$  is a real  $n \times n$  matrix and  $b$  is a real  $n$ -vector. We decompose  $A$  into its symmetric and skew-symmetric parts,

$$(2.2) \quad A = P + Q$$

where  $P$  is the symmetric part of  $A$  and  $Q$  is the skew-symmetric part, i.e.

$$P = \frac{1}{2} (A + A^T), \quad Q = \frac{1}{2} (A - A^T).$$

We assume that  $P$  is positive definite and hence the system (2.1) has a unique solution.

We will begin by considering the special case of equation (1.1) where  $P$  is the identity matrix. In section 5 we will show that the method can treat those cases where the system

$$Pz = c$$

can be easily solved. This is the same situation considered by Concus and Golub (1976) and Widlund (1978).

The standard conjugate gradient method (Hestenes and Stiefel (1952)) is used to solve linear systems such as (2.1) when the matrix  $A$  is symmetric positive definite. The method may be described as

$$(2.3) \quad \begin{aligned} a) \quad & x^{k+1} = x^k + \gamma_k p^k \\ b) \quad & r^{k+1} = r^k - \gamma_k A p^k \\ c) \quad & p^{k+1} = r^{k+1} + \delta_k p^k \end{aligned}$$

where the parameters  $\gamma_k$  and  $\delta_{k-1}$  are determined so as to minimize  $(r^{k+1}, A^{-1}r^{k+1})$  given  $r^k$  and  $p^{k-1}$ . The vector  $r^k$  is the residual  $b - Ax^k$  and so (2.3b) is a consequence of (2.3a). The values of  $\gamma_k$  and  $\delta_k$  are

$$(2.4) \quad \gamma_k = |r^k|^2 / (p^k, Ap^k)$$

$$\delta_k = |r^{k+1}|^2 / |r^k|^2 .$$

By analogy with the conjugate gradient method for positive definite systems we consider the following iterative scheme for the system (1.1)

$$(2.5) \quad \begin{aligned} a) \quad w^{k+1} &= w^k + \alpha_k p^k \\ b) \quad r^{k+1} &= r^k - \alpha_k (I+S)p^k \\ c) \quad p^{k+1} &= r^{k+1} - \beta_k p^k . \end{aligned}$$

The parameters  $\alpha_k$  and  $\beta_{k-1}$  are to be chosen to minimize  $|r^{k+1}|^2$  given  $r^k$  and  $p^{k-1}$ . Equation (2.5b) is a consequence of the definition of the residual vector,  $r^k = b - (I+S)w^k$ , and (2.5a).

One obtains a steepest descent method by setting all  $\beta_k = 0$ , so  $p^k = r^k$ , and then  $\alpha_k = \alpha'_k$ , where

$$\alpha'_k = |r^k|^2 / (|r^k|^2 + |Sp^k|^2)$$

minimizes  $|r^{k+1}|^2$ .

To derive our method we choose  $\alpha_k$  to minimize  $|r^{k+1}|^2$  given  $p^k$ . We have

$$(2.6) \quad |r^{k+1}|^2 = |r^k|^2 - 2\alpha_k (r^k, (I+S)p^k) + \alpha_k^2 |(I+S)p^k|^2$$

and so

$$(2.7) \quad \alpha_k = (r^k, (I+S)p^k) / (|p^k|^2 + |Sp^k|^2)$$

is the optimal value of  $\alpha_k$ . The first consequence of (2.7) is that

$$(2.8) \quad (r^{k+1}, (I+S)p^k) = 0$$

by (2.5b), and secondly, using (2.8) with (2.5c)

$$\begin{aligned} (r^k, (I+S)p^k) &= (r^k, (I+S)(r^k - \beta_{k-1}p^{k-1})) \\ &= (r^k, (I+S)r^k) - \beta_{k-1}(r^k, (I+S)p^{k-1}) \\ &= |r^k|^2 . \end{aligned}$$

Hence

$$(2.9) \quad \alpha_k = |r^k|^2 / (|p^k|^2 + |Sp^k|^2) .$$

The relation (2.6) then becomes

$$(2.10) \quad |r^{k+1}|^2 = |r^k|^2 (1 - \alpha_k)$$

and since  $|r^{k+1}| < |r^k|$  we have

$$0 < \alpha_k < 1 .$$

From (2.8) and (2.5b) it also follows that

$$(2.11) \quad |r^{k+1}|^2 = (r^{k+1}, r^k) .$$

We now determine  $\beta_{k-1}$  given  $r^k$  and  $p^{k-1}$ . We see from (2.10) and (2.9) that, given  $r^k$ ,  $|r^{k+1}|^2$  is minimized when  $\alpha_k$  is maximized and this requires that  $|p^k|^2 + |Sp^k|^2$  be minimized. We have

$$\begin{aligned} |p^k|^2 + |Sp^k|^2 &= ((I+S)p^k, (I+S)p^k) \\ &= |(I+S)r^k|^2 - 2\beta_{k-1}((I+S)r^k, (I+S)p^{k-1}) \\ &\quad + \beta_{k-1}^2 |(I+S)p^{k-1}|^2 , \end{aligned}$$

and hence

$$(2.12) \quad \beta_{k-1} = ((I+S)r^k, (I+S)p^{k-1}) / (|p^{k-1}|^2 + |Sp^{k-1}|^2) .$$

From (2.12) and (2.5c) we have

$$(2.13) \quad ((I+S)p^k, (I+S)p^{k-1}) = 0 ,$$

and by subtracting  $\alpha_{k-1}$  times (2.13) from (2.8) with  $k-1$  replacing  $k$  we obtain

$$(2.14) \quad (r^{k+1}, (I+S)p^{k-1}) = 0 .$$

We now wish to obtain an expression for  $\beta_k$  which is simpler than (2.12). We have

$$\begin{aligned} ((I+S)r^{k+1}, (I+S)p^k) &= (Sr^{k+1}, (I+S)p^k) \text{ by (2.8)} \\ &= \frac{1}{\alpha_k} (Sr^{k+1}, r^k - r^{k+1}) \text{ by (2.5b)} \\ &= \frac{1}{\alpha_k} (Sr^{k+1}, r^k) . \end{aligned}$$

Hence

$$\beta_k = (Sr^{k+1}, r^k) / |r^k|^2 .$$

However,

$$\begin{aligned} (Sr^{k+1}, r^k) &= -(r^{k+1}, Sr^k) \\ &= -(r^{k+1}, (I+S)r^k) + |r^{k+1}|^2 \text{ by (2.11)} \\ &= -(r^{k+1}, (I+S)(p^k + \beta_{k-1}p^{k-1})) + |r^{k+1}|^2 \text{ by (2.5c)} \\ &= |r^{k+1}|^2 \text{ by (2.8) and (2.13)} . \end{aligned}$$

So

$$(2.15) \quad \beta_k = |r^{k+1}|^2 / |r^k|^2 = 1 - \alpha_k \text{ by (2.10)} .$$

We now summarize the algorithm.

$$\begin{aligned} (2.16) \quad \text{a)} \quad w^{k+1} &= w^k + \alpha_k p^k \\ \text{b)} \quad r^{k+1} &= r^k - \alpha_k (I+S)p^k \\ \text{c)} \quad p^{k+1} &= r^{k+1} - \beta_k p^k \end{aligned}$$

with

$$\text{d)} \quad \alpha_k = |r^k|^2 / (|p^k|^2 + |Sp^k|^2)$$

and  $\beta_k = 1 - \alpha_k$ .

For initial vectors we take  $w^0$  arbitrary,  $r^0 = p^0 = b - (I+S)w^0$ .

### 3. Orthogonality Relations.

The purpose of this section is to prove orthogonality relations for the vectors generated by the above algorithm. The main result is contained in the following theorem.

#### Theorem 3.1

For the algorithm (2.16)

$$(3.1) \quad ((I+S)p^i, (I+S)p^j) = 0$$

for  $i \neq j$ .

#### Proof

We begin by obtaining a three term recurrence relation for the vectors  $p^k$ . By eliminating the vectors  $r^k$  from (2.16b) and (2.16c) we obtain

$$(3.2) \quad p^{k+1} + \alpha_k S p^k - \beta_{k-1} p^{k-1} = 0$$

for  $k > 0$  where  $p^{-1} = 0$ .

First we show that (3.1) is true for  $i = 1, j = 0$ . This is immediate from (3.2) and the skew-symmetry of  $S$

$$\begin{aligned} ((I+S)p^1, (I+S)p^0) &= -\alpha_0 ((I+S)Sp^0, (I+S)p^0) \\ &= 0 . \end{aligned}$$

Now assume that (3.1) is true for  $k > i > j > 0$ , we will show that it is also true for  $k+1 = i > j > 0$ . If  $i = k+1$  and  $j = k$ , (3.1) is true by (2.13). If  $i = k+1$  and  $j = k-1$ , we have by (3.2)

$$\begin{aligned} (3.3) \quad &((I+S)p^{k+1}, (I+S)p^{k-1}) \\ &= -\alpha_k ((I+S)Sp^k, (I+S)p^{k-1}) + \beta_{k-1} |(I+S)p^{k-1}|^2 . \end{aligned}$$

Consider the first expression on the right-hand side of (3.3).

$$\begin{aligned} &((I+S)Sp^k, (I+S)p^{k-1}) \\ &= -((I+S)p^k, (I+S)Sp^{k-1}) \end{aligned}$$

$$= -((I+S)p^k, (I+S)(-p^k + \beta_{k-2}p^{k-2}))/\alpha^{k-1}$$

$$= |(I+S)p^k|^2 / \alpha_{k-1} .$$

Thus (3.3) is equal to

$$- \frac{\alpha_k}{\alpha_{k-1}} |(I+S)p^k|^2 + \beta_{k-1} |(I+S)p^{k-1}|^2 ,$$

and by the expressions (2.16d) and (2.15) this is equal to

$$- \frac{|r^k|^2}{|r^{k-1}|^2} |(I+S)p^{k-1}|^2 + \frac{|r^k|^2}{|r^{k-1}|^2} |(I+S)p^{k-1}|^2$$

$$= 0 .$$

Now for  $i = k+1$  and  $j < k-1$  the relation (3.2) follows easily.

$$((I+S)p^{k+1}, (I+S)p^j)$$

$$= -\alpha_k ((I+S)p^k, (I+S)p^j) + \beta_{k-1} ((I+S)p^k, (I+S)p^j)$$

$$= \alpha_k ((I+S)p^k, (I+S)p^j)$$

$$= \frac{\alpha_k}{\alpha_j} ((I+S)p^k, (I+S)(p^j - \beta_{j-1}p^{j-1}))$$

$$= 0 .$$

Thus the theorem is proved.

It follows that the vectors  $p^k$  are linearly independent, as long as they are non-zero, and thus the algorithm must converge in at most  $N$  steps, where  $N$  is the dimension of the vector space.

Other orthogonality relations are given in the next theorem.

### Theorem 3.2

For the algorithm (2.16)

$$(3.4) \quad (r^k, (I+S)r^j) = 0 \text{ for } j < k$$

$$(3.5) \quad (r^k, (I+S)p^j) = 0 \text{ for } j < k$$

$$(3.6) \quad (p^{k+1}, p^k) = 0 \text{ for } 0 \leq k$$

$$(3.7) \quad (r^k, r^j) = (r^k, r^0) \text{ for } j \leq k .$$

Proof of (3.4)

For  $k = 1$  and  $j = 0$  (3.4) follows from (2.8) since  $p^0 = r^0$ . When  $j = k-1 > 0$ ,  $(r^k, (I+S)r^{k-1}) = (r^k, (I+S)(p^{k-1} + \beta_{k-2}p^{k-2})) = 0$  by (2.8) and (2.14). For  $j < k-1$  the result follows from Theorem 3.1 by induction, we have

$$\begin{aligned}(r^k, (I+S)r^j) &= (r^{k-1}, (I+S)r^j) - \alpha_{k-1}((I+S)p^{k-1}, (I+S)r^j) \\ &= -\alpha_{k-1}((I+S)p^{k-1}, (I+S)(p^j + \beta_{j-1}p^{j-1})) \\ &= 0.\end{aligned}$$

Proof of (3.5)

By repeated use of (2.16c) and (3.4)

$$\begin{aligned}(r^k, (I+S)p^j) &= (r^k, (I+S)r^j) - \beta_{j-1}(r^k, (I+S)p^{j-1}) \\ &= -\beta_{j-1}(r^k, (I+S)p^{j-1}) \\ &= (-)^j \prod_{l=0}^{j-1} \beta_l(r^k, (I+S)p^0) = 0\end{aligned}$$

by (3.4) since  $p^0 = r^0$ .

Proof of (3.6).

By (3.2)

$$\begin{aligned}(p^{k+1}, p^k) &= \beta_{k-1}(p^{k-1}, p^k) \\ &= \prod_{l=0}^{k-1} \beta_l(p^1, p^0).\end{aligned}$$

Now by (2.16c) and (2.16b)

$$\begin{aligned}(p^1, p^0) &= (r^1, p^0) - \beta_0 |p^0|^2 \\ &= (r^0, p^0) - \alpha_0 ((I+S)p^0, p^0) - \beta_0 |p^0|^2 \\ &= (r^0, p^0) - (\alpha_0 + \beta_0) |p^0|^2 = 0,\end{aligned}$$

since  $p^0 = r^0$  and  $\alpha_0 + \beta_0 = 1$ .

Proof of (3.7).

By (2.16b) for  $j < k$

$$\begin{aligned}(r^k, r^j) &= (r^k, r^{j-1}) - \alpha_{j-1}(r^k, (I+S)p^{j-1}) \\&= (r^k, r^{j-1}) \quad \text{by (3.5)} \\&= (r^k, r^0), \quad \text{by repetition.}\end{aligned}$$

This proves all the assertions of Theorem 3.2.

The relation (3.7) has the geometric interpretation that  $r^k$  is on the sphere of radius  $\frac{1}{2} |r^j|$  centered at  $\frac{1}{2} r^j$  for each  $j$  less than  $k$ . This again demonstrates the finite termination property of the method since  $N$  distinct spheres through the origin in  $N$ -space can have only the origin as a common point.

#### 4. The Rate of Convergence

Considering the method (2.16) as an iterative method for solving (1.1), it is natural to estimate the rate of convergence of the method in terms of the spectral radius of  $S$ . For linear systems with a large number of unknowns, such as arise from numerical approximations to partial differential equations, the finite termination property is of little interest compared with the rate of convergence.

Our first convergence results are stated in the following theorem.

##### Theorem 4.1

For the method (2.16) the following estimates hold.

$$(4.1) \quad |r^{k+1}|/|r^k| < \frac{\Lambda}{\sqrt{1 + \Lambda^2}}$$

$$(4.2) \quad |r^{k+2}|/|r^k| < \left( \frac{\Lambda}{\sqrt{2 + \Lambda^2}} \right)^2$$

where  $\Lambda$  is the spectral radius of  $S$ .

##### Proof

We begin with

$$\begin{aligned} |r^{k+1}|^2 &= (r^{k+1}, r^k) && \text{by (3.7)} \\ &= -(r^{k+1}, Sr^k) && \text{by (3.4)} \\ &= \alpha_k ((I+S)r^k, Sr^k) && \text{by (2.16b)} \\ &= \alpha_k ((I+S)r^k, Sr^k) - \alpha_k \beta_{k-1} ((I+S)r^{k-1}, Sr^k) && \text{by (2.16c)} \\ &= \alpha_k |Sr^k|^2 - \frac{\alpha_k \beta_{k-1}}{\alpha_{k-1}} (r^{k-1} - r^k, Sr^k) && \text{by (2.16b)} \\ &= \alpha_k |Sr^k|^2 - \frac{\alpha_k \beta_{k-1}}{\alpha_{k-1}} |r^k|^2 && \text{by (3.4) and (3.7)} . \end{aligned}$$

This gives the estimate

$$(4.3) \quad \frac{\beta_{k-1}}{\alpha_{k-1}} + \frac{\beta_k}{\alpha_k} = \frac{|Sr^k|^2}{|r^k|^2} < \Lambda^2$$

$$\text{since } \beta_k = |r^{k+1}|^2 / |r^k|^2.$$

The estimate (4.1) follows from (4.3), since

$$\frac{\beta_k}{\alpha_k} < \Lambda^2$$

and  $\alpha_k = 1 - \beta_k$ . So

$$|r^{k+1}|^2 / |r^k|^2 = \beta_k < \frac{\Lambda^2}{1 + \Lambda^2}$$

which is (4.1).

The estimate (4.2) is obtained by finding the maximum of the product

$$\begin{aligned} \beta_{k+1}\beta_k &= (1 - \alpha_{k+1})(1 - \alpha_k) \text{ subject only to} \\ (4.4) \quad \frac{1}{\alpha_{k+1}} + \frac{1}{\alpha_k} &< \Lambda^2 + 2 \end{aligned}$$

which is equivalent to (4.3). The maximum of  $(1 - \alpha_{k+1})(1 - \alpha_k)$  obviously occurs when equality holds in (4.4). Thus

$$\begin{aligned} (1 - \alpha_{k+1})(1 - \alpha_k) &= 1 - \alpha_{k+1} - \alpha_k + \alpha_{k+1}\alpha_k \\ &= 1 - \alpha_{k+1}\alpha_k \left( \frac{1}{\alpha_k} + \frac{1}{\alpha_{k+1}} - 1 \right) \\ &= 1 - \alpha_{k+1}\alpha_k (\Lambda^2 + 1) \end{aligned}$$

and this quantity is maximized subject to (4.4) when

$$\alpha_{k+1}^{-1} = \alpha_k^{-1} = \frac{1}{2}(\Lambda^2 + 2). \text{ Hence}$$

$$\begin{aligned} \beta_{k+1}\beta_k &< 1 - (\Lambda^2 + 1)/(\frac{1}{2}\Lambda^2 + 1)^2 \\ &= \left( \frac{\Lambda^2}{2 + \Lambda^2} \right)^2 \end{aligned}$$

which is (4.2). This proves Theorem 4.1.

More general results can be obtained by a method similar to that of Widlund (1978).

#### Theorem 4.2

For the method (2.16), for  $k$  even or  $k = 1$ ,

$$(4.5) \quad |r^k|/|r^0| < 2\rho^k/(1 + \rho^{2k})$$

and for  $k$  odd

$$(4.6) \quad |r^k|/|r^0| < 2\rho^k/(1 - \rho^{2k})$$

and

$$(4.7) \quad |r^k|/|r^0| < 4\rho^k/((1 + \rho^2)(1 + \rho^{2(k-1)})) ,$$

where  $\rho = \Lambda/\sqrt{1 + \Lambda^2} + 1$ . Note that (4.6) is a better estimate than (4.7) when  $\Lambda$  is large and  $k$  is large, (4.7) is better when  $\Lambda$  or  $k$  is small. Note that (4.5) for  $k = 1$  and  $k = 2$  gives the same result as Theorem 4.1.

#### Proof of Theorem 4.2

By (3.4) and (3.7)

$$(r^k, r^k) = (r^k, z + r^0)$$

where  $z$  is in the span of  $(I+S)r^0, \dots, (I+S)r^{k-1}$ . Now by (2.16)

$z = P_k(I+S)r^0$  where  $P_k(\lambda)$  is a polynomial of degree  $k$  with  $P_k(0) = 0$ .

Thus

$$|r^k|^2 = (r^k, Q_k(I+S)r^0)$$

where  $Q_k(\lambda)$  is a polynomial of degree  $k$  and  $Q_k(0) = 1$ . By the spectral mapping theorem

$$|r^k| < \min_{Q_k(0)=1} \max_{\mu \in \sigma(S)} |Q_k(1+\mu)| |r^0|$$

where  $\sigma(S)$  is the spectrum of  $S$ . Since  $S$  is skew-symmetric  $\sigma(S)$  is contained in the imaginary axis with  $-1 < \mu/i\Lambda < 1$ . The minimum is taken over all polynomials  $Q_k(\lambda)$  of degree  $k$  with  $Q_k(0) = 1$ . As does Widlund (1978), we take the particular polynomial

$$Q_k(1+\mu) = \frac{T_k(\mu/i\Lambda)}{T_k(-1/i\Lambda)}$$

where  $T_k(\lambda) = \cosh(k \cosh^{-1} \lambda)$ . We have  $|T_k(\mu/i\Lambda)| = |\cos(k \cos^{-1}(\mu/i\Lambda))| < 1$  and

$$\begin{aligned} T_k(-1/i\Lambda) &= \cosh(k \cosh^{-1}(-1/i\Lambda)) = \cosh(k \log(\sqrt{1 + \Lambda^2} + 1)\Lambda^{-1}) \\ &= \frac{1}{2} ((\sqrt{1 + \Lambda^2} + 1)^k + (\sqrt{1 + \Lambda^2} - 1)^k)\Lambda^{-k} \\ &= \frac{1}{2} (1 + \rho^{2k})\rho^{-k} \end{aligned}$$

for  $k$  odd and

$$\begin{aligned} |T_k(-1/i\Lambda)| &= |\sinh(k \log(\sqrt{1 + \Lambda^2} + 1)\Lambda^{-1})| \\ &= \frac{1}{2} ((\sqrt{1 + \Lambda^2} + 1)^k - (\sqrt{1 + \Lambda^2} - 1)^k)\Lambda^{-k} \\ &= \frac{1}{2} (1 - \rho^{2k})\rho^{-k} \end{aligned}$$

for  $k$  even.

This proves (4.5) for  $k \neq 1$  and (4.6). Inequality (4.5) for  $k = 1$  is just (4.1), and (4.7) follows from (4.5) and (4.1). This proves Theorem 4.2.

The above theorems give results on the error vectors  $e^k$  since  $r^k = (I+S)e^k$  and so

$$(4.8) \quad |e^k| < |r^k| < \sqrt{1 + \Lambda^2} |e^k|.$$

## 5. The case with general symmetric part

In this section we discuss the more general case when  $P$ , the symmetric part of  $A$ , is not the identity. In this case the system

$$(5.1) \quad Ay = (P + Q)y = c$$

can be transformed to one of the form (1.1) by setting

$$(5.2) \quad x = P^{1/2}y, \quad b = P^{-1/2}c, \quad s = P^{-1/2}QP^{-1/2}.$$

The algorithm (2.16) can then be used on the resulting system. It is, of course, more convenient to work with the original matrices  $P$  and  $Q$  than with  $S$ , thus we consider the system (2.16) in the original variables. We have

$$(5.3) \quad \begin{aligned} a) \quad w^{k+1} &= w^k + \alpha_k p^k \\ b) \quad r^{k+1} &= r^k - \alpha_k A p^k \\ c) \quad p^{k+1} &= P^{-1}r^{k+1} - \beta_k p^k \end{aligned}$$

where

$$(5.3) \quad d) \quad \alpha_k = (r^k, P^{-1}r^k) / (A p^k, P^{-1}A p^k)$$

and  $\beta_k = 1 - \alpha_k$ .

The vector  $w^0$  is arbitrary, and  $p^0 = P^{-1}r^0$  where  $r^0 = c - Aw^0$ . The vectors  $w^k$  converge to  $y$  the solution of (5.1). Because of the necessity of computing  $P^{-1}r^{k+1}$  and  $P^{-1}Ap^k$  the method is applicable in practice only when the solution of

$$Pz = d$$

can be obtained easily. This restriction applies also to the method of Concus and Golub (1976) and Widlund (1978). Note that  $P^{-1}r^{k+1}$  can be obtained by

$$P^{-1}r^{k+1} = P^{-1}r^k - \alpha_k P^{-1}Ap^k$$

so only one inversion of  $P$  is required per iteration step.

The orthogonality results for the algorithm (5.3) are easily deduced from the results for (2.16). We state these results for completeness.

Theorem 5.1

For the algorithm (5.3) the following relations hold.

$$\begin{aligned} (\mathbf{A}\mathbf{p}^i, \mathbf{P}^{-1}\mathbf{A}\mathbf{p}^j) &= 0 \quad \text{for } i \neq j \\ (\mathbf{P}^{-1}\mathbf{r}^k, \mathbf{A}\mathbf{P}^{-1}\mathbf{r}^j) &= 0 \quad \text{for } j < k \\ (\mathbf{P}^{-1}\mathbf{r}^k, \mathbf{A}\mathbf{p}^j) &= 0 \quad \text{for } j < k \\ (\mathbf{p}^{k+1}, \mathbf{P}\mathbf{p}^k) &= 0 \quad \text{for } k > 0 \\ (\mathbf{r}^k, \mathbf{P}^{-1}\mathbf{r}^j) &= (\mathbf{r}^k, \mathbf{P}^{-1}\mathbf{r}^0) \quad \text{for } j < k . \end{aligned}$$

Theorem 5.2

For the algorithm (5.3) the following estimates holds.

$$|\mathbf{r}^k|_{\mathbf{P}^{-1}} / |\mathbf{r}^0|_{\mathbf{P}^{-1}} \leq 2\rho^k / (1 + \rho^{2k}) \quad \text{for } k \text{ even or } k = 1$$

$$|\mathbf{r}^k|_{\mathbf{P}^{-1}} / |\mathbf{r}^0|_{\mathbf{P}^{-1}} \leq 2\rho^k / (1 - \rho^{2k}) \quad \text{for } k \text{ odd}$$

and

$$|\mathbf{r}^k|_{\mathbf{P}^{-1}} / |\mathbf{r}^0|_{\mathbf{P}^{-1}} \leq 4\rho^k / ((1 + \rho^2)(1 + \rho^{2(k-1)})) \quad \text{for } k \text{ odd}$$

where  $|\mathbf{r}^k|_{\mathbf{P}^{-1}}^2 = (\mathbf{r}^k, \mathbf{P}^{-1}\mathbf{r}^k)$ ,  $\rho = \Lambda / (\sqrt{1 + \Lambda^2} + 1)$  and  $\Lambda$  is the spectral radius of  $\mathbf{P}^{-1/2}\mathbf{Q}\mathbf{P}^{-1/2}$ .

The proofs of these results follow from Theorems 3.1, 3.2, and 4.2.

Corresponding to (4.8) we have

$$(5.4) \quad (\mathbf{e}^k, \mathbf{P}\mathbf{e}^k) \leq (\mathbf{r}^k, \mathbf{P}^{-1}\mathbf{r}^k) \leq \sqrt{1 + \Lambda^2} (\mathbf{e}^k, \mathbf{P}\mathbf{e}^k)$$

which in conjunction with Theorem 5.2 gives estimates for the error.

## 6. Numerical Experiments.

The performance of the general algorithm (5.3) depends significantly on the means of inverting the positive definite matrix  $P$ . Therefore, to test the algorithm it seemed best to study only the basic algorithm (2.16) i.e. with  $P$  being the identity. A FORTRAN computer code was written which implemented the algorithm (2.16). The program was run on the UNIVAC 1100 computer at the Madison Academic Computing Center using single precision arithmetic which has about seven digits of accuracy.

The computer code also implemented the algorithm of Concus and Golub (1976) and Widlund (1978). We will refer to this algorithm as the CGW algorithm. The CGW algorithm and (2.16) were run simultaneously but independently.

The matrices  $S$  used in the experiments were banded skew-symmetric matrices whose non-zero elements were generated randomly. For  $0 < i - j \leq m$ ,  $S_{ij}$  was a randomly generated floating point number in the interval  $[-\delta, \delta]$  for some number  $\delta$ ,  $0 < \delta \leq 1$ , also  $S_{ji} = -S_{ij}$ , otherwise  $S_{ij}$  was zero. For all the numerical experiments  $x_i^0 = 0$ , with the exact solution being  $x_i = 1$ . The algorithm (2.16) was stopped when  $|r^k|/|r^0| < 10^{-5}$ .

Table I displays the results from several experiments.  $N$  is the number of unknowns,  $m$  gives the size of the band and  $\delta$  is the range of random numbers for  $S$  as described above.  $\Lambda$  is the spectral radius for  $S$  as computed by EISPACK routines (Smith et al. 1976). The fifth column gives  $I$ , the number of iterations required for convergence, and the sixth and seventh columns give the values of  $|e^I|/|e^0|$  for both the algorithm (2.16) and the CGW algorithm. The eighth column contains the quantity appearing on the right-hand side of the estimate (4.5), for  $k = I$ .

It was found that the two algorithms converged at about the same rate with (2.16) having slightly smaller values for the norms of the error and residual vectors. The similarity is not surprising since the error estimate (4.5), using (4.8), is similar to that given by Widlund (1978) for the CGW algorithm. For the algorithm (2.16) it was found in all the experiments that the norm of the error decreased monotonically. This was different than the CGW algorithm for which the norm of the error usually did not decrease monotonically for the first several iterates. However, for the CGW algorithm the even and odd iterates do give monotonically decreasing values for the norm of the error, Widlund (1978).

The estimate (4.5) is seen to give a good approximation of the behavior of the algorithm and can be used to give a good estimate of the number of iterations required for convergence.

Table I: Results from Numerical Experiments

N	m	$\delta$	$\Lambda$	I	error	error (CGW)	estimate (4.5)
20	3	0.2	0.48	8	.62E-5	.63E-5	.14E-4
		0.6	1.62	15	.40E-5	.41E-5	.32E-3
		1.0	2.30	17	.60E-5	.61E-5	.15E-2
20	5	0.2	0.59	9	.74E-5	.76E-5	.16E-4
		0.6	1.72	16	.14E-4	.14E-4	.29E-3
		1.0	2.74	18	.87E-5	.90E-5	.32E-2
40	3	0.2	0.48	8	.90E-5	.93E-5	.14E-4
		0.6	1.34	17	.80E-5	.86E-5	.16E-4
		1.0	2.47	24	.78E-5	.83E-5	.15E-3
40	5	0.2	0.64	10	.34E-5	.34E-5	.89E-5
		0.6	1.79	21	.58E-5	.59E-5	.27E-4
		1.0	3.12	29	.13E-4	.16E-4	.21E-3
80	3	0.2	0.54	9	.47E-5	.48E-5	.85E-5
		0.6	1.43	18	.70E-5	.76E-5	.16E-4
		1.0	2.45	28	.78E-5	.83E-5	.29E-4
80	5	0.2	0.64	10	.61E-5	.63E-5	.89E-5
		0.6	1.86	23	.69E-5	.76E-5	.14E-4
		1.0	3.64	37	.14E-4	.16E-4	.87E-4

#### REFERENCES

- [1] P. Concus and G. H. Golub (1976), A generalized conjugate gradient method for nonsymmetric systems of linear equations. Proc. Second Internat. Symp. on Computing Methods in Applied Sciences and Engrg., IRIA (Paris, Dec. 1975) Lect. Notes in Econ. and Math. Systems, vol. 134, R. Glowinski and J. L. Lions, eds., Springer-Verlag, Berlin.
- [2] L. A. Hageman, F. T. Luk, and D. M. Young (1980), On the equivalence of certain iterative acceleration methods, SIAM J. Numer. Anal., 17, pp. 852-873.
- [3] M. R. Hestenes and E. Stiefel (1952), Method of conjugate gradients for solving linear systems, J. Res. Nat. Bur. Standards, 49, pp. 409-436.
- [4] B. T. Smith, J. M. Boyle, J. J. Dongarra, B. S. Garbow, Y. Ikebe, V. C. Klema, and C. B. Moler (1976), Matrix Eigensystem Routines - EISPACK Guide, Second Edition, Lecture Notes in Computer Science 6, Springer-Verlag, Berlin.
- [5] O. Widlund (1978), A Lanczos method for a class of nonsymmetric systems of linear equations, SIAM J. Numer. Anal., 15, pp. 801-812.

JCS/jvs

**SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)**

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER  #2290	2. GOVT ACCESSION NO.  <i>AD-A110 311</i>	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle)  A Generalized Conjugate Gradient Method for Non-Symmetric Systems of Linear Equations	5. TYPE OF REPORT & PERIOD COVERED  Summary Report - no specific reporting period	
7. AUTHOR(s)  John C. Strikwerda	6. PERFORMING ORG. REPORT NUMBER  NASI-15810 and NASI-16394 DAAG29-80-C-0041 MCS-7927062	
9. PERFORMING ORGANIZATION NAME AND ADDRESS  Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS  Work Unit Number 3 - Numerical Analysis and Computer Science	
11. CONTROLLING OFFICE NAME AND ADDRESS  (see Item 18 below)	12. REPORT DATE  October 1981	
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	13. NUMBER OF PAGES  19	
	15. SECURITY CLASS. (of this report)  UNCLASSIFIED	
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE	
16. DISTRIBUTION STATEMENT (of this Report)  Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES  U. S. Army Research Office P. O. Box 12211 Research Triangle Park North Carolina 27709	National Science Foundation Washington, DC 20550	National Aeronautics and Space Administration Washington, DC 20546
19. KEY WORDS (Continue on reverse side if necessary and identify by block number)  Conjugate-Gradient Method, Non-symmetric Systems		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number)  A new iterative method is presented for solving non-symmetric linear systems of equations. The method requires that the symmetric part of the matrix of the linear system be positive definite, and the method is efficient only if the symmetric part is easily invertible. The method is modeled on the conjugate gradient method for symmetric positive definite systems and has the finite termination property. The results from several numerical experiments are presented and compared with a similar method proposed by Concus, Golub, and Widlund.		

**DD FORM 1473 EDITION OF 1 NOV 68 IS OBSOLETE**

**UNCLASSIFIED**

**SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)**

2-8